DOCUMENT RESUME

ED 074 128

TM 002 491

AUTHOR Bergsten, Jane Williams

TITLE The Effects of Cluster Sampling in the Norming of an Achievement Test Battery.

PUB DATE Feb 73

NOTE 8p.; Paper presented at annual meeting of the American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS *Achievement Tests; *Cluster Analysis; Grade 4; *Grade Equivalent Scores; *Norms; Speeches; Technical Reports; *Test Results

IDENTIFIERS *Iowa Tests of Basic Skills

ABSTRACT

Using the grade equivalent composite scores on the Iowa Tests of Basic Skills of Iowa fourth grade public school pupils who took the tests in January 1970, a study was made to determine the relative precision with which an estimate could be made of the individual percentile norms from different types of cluster sample designs. Five scores ranging from the 14th to the 93rd percentiles were selected, and the proportions below these five scores became the proportions to be estimated. The variances of the estimates of these five proportions were computed for over 20 different sample designs; results from seven sample designs are presented. Using the error variances that were computed for each of the seven sample designs, the ratio of the error variance based on a cluster sample to the error variance based on a simple random sample of pupils was determined. (DB)

FILMED FROM BEST AVAILABLE COPY

AERA, 1973 Annual Meeting
February, 1973

ED 074128

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

THE EFFECTS OF CLUSTER SAMPLING IN THE NORMING OF AN ACHIEVEMENT TEST BATTERY

by Jane Williams Bergsten

TM 002 491

In constructing individual pupil percentile norms for test scores one would typically test a sample of pupils rather than an entire population of pupils. In selecting and testing a sample of pupils it is usually administratively easier and less expensive per pupil to test whole classes, whole buildings or whole systems of pupils than it is to sample and test pupils individually. However, assuming that a given number of pupils are to be tested, the use of any of these types of cluster samples could be expected to decrease the precision of the estimates which could be made. The objective of this study was to determine the relative precision with which one could estimate the individual pupil percentile norms for pupil achievement test scores of the children in a particular grade from different types of cluster sample designs.

The population studied consisted of the grade equivalent composite scores on the Iowa Tests of Basic Skills made by Iowa fourth grade public school pupils who took the battery of tests in January of 1970. This consisted of scores for about 95% of all fourth grade public school pupils in the state or approximately 45,000 fourth graders in 1,800 classes in 1,000 buildings in

400 systems. The entire population of scores was included in the analysis. Thus, the values of the population proportions to be estimated were actually known, as were the sizes of the various clusters studied.

Since from a practical point of view it was not feasible to attempt to estimate proportions below every score, five scores ranging from the 14th to the 93rd percentiles were selected, and the proportions below these five scores became the proportions to be estimated. The variances of the estimates of these five proportions were computed for over 20 different sample designs. The results from seven sample designs will be presented here. These include random samples of entire classes, of entire buildings, and of entire school systems of fourth grade pupils. Using the error variances that were computed for each of the seven sample designs, the ratio of the error variance based on a cluster sample to the error variance based on a simple random sample of pupils was determined. These ratios were measures of the relative inefficiency of the cluster sample design and are commonly referred to as measures of design effects or DEFFs. DEFFs were used to compare the relative precision of estimates based on different sample designs. In addition, DEFFs were used to compute the intraclass correlation coefficients for classes, buildings and systems. The intraclass correlation coefficient provides a measure of the within cluster homogeneity.

From the handout you can see that a design involving the selection of whole classes of fourth grade pupils yielded error variances roughly four

times those of a simple random sample while the selection of fourth graders by buildings yielded variances about six or seven times those of a simple random sample. When whole systems of fourth graders were selected, variances up to 56 times those of a simple random sample were obtained.

There is another way of looking at these figures. In order to obtain the same error variance as would be obtained from a simple random sample, it would be necessary to test about four times as many pupils if they were selected by classes, about six or seven times as many pupils if they were selected by buildings, and up to 56 times as many pupils if they were selected by systems.

By partitioning the systems into four strata according to the total K-12 enrollment and selecting clusters separately from within each stratum, the error variances were reduced. As compared to a comparable unstratified design, a stratified sample of classes produced a reduction in error variance of about 10%. In the case of sampling buildings, a reduction in error variance of about 15% occurred because of stratification, and in the case of sampling systems, there was a reduction of about 45%.

One of the reasons for the large error variances in the case of sampling whole systems of fourth graders is that some of the systems had very large enrollment, therefore a large number of fourth graders. If a "mixed" design were used wherein buildings were selected from the stratum containing the largest systems, and systems were selected from each of the other three

strata, the error variances were reduced to about one sixth of those for an unstratified design sampling whole systems of fourth graders.

Consistently for each of the seven sample designs the DEFFs obtained for estimates of proportions in the lower part of the distribution were greater than for estimates in the upper part, indicating that classes, buildings and school systems are more homogeneous with respect to the characteristic of having pupils in one of the lower score groups than in one of the upper score groups. In other words, lower scoring pupils tend to be more highly concentrated in certain classes, buildings and systems than higher scoring pupils who tend to be more evenly spread throughout the population.

Although error variances were found to be smaller for sampling whole classes of fourth grade pupils than for sampling whole buildings, and smaller for sampling whole buildings of fourth graders than for sampling whole systems, the relationships among the intraclass correlations were just the reverse, being greatest for classes and smallest for systems. In other words, classes of fourth grade pupils are more homogeneous than buildings of fourth graders, which in turn are more homogeneous than systems of fourth graders. The apparent inconsistency in the relationships is due to the differences in average cluster size, which was about 100 fourth graders per system, 50 per building and 25 per class. Since the relationship between DEFF, ρ, and average cluster size is

$$\text{DEFF} \doteq 1 + \rho(\overline{N} - 1)$$

where

$\overline{N}$ = the average size of cluster,

it is apparent that when $\overline{N}$ is doubled while ρ is decreased only slightly, DEFF can increase.

Conclusions

Judging from the results of this investigation it is not safe to assume that the homogeneity within classes, buildings or school systems is small enough to be ignored. Indeed, the homogeneity is sufficient to require much larger samples of pupils if they are to be selected by classes, buildings or school systems. However, since pupils can usually be tested much more cheaply by testing entire classes than they could be individually, a cluster sample will usually be the best choice economically, even when it is necessary to include many more pupils than would be needed if a simple random sample were to be used.

With respect to generalizing the findings, we can make some educated guesses about the way in which the results of this study would compare to the results that would be obtained for other variables, grades and states.

First, with respect to the 11 subtests that are covered by the Iowa Tests of Basic Skills which include tests in vocabulary, reading, language, work-study and arithmetic skills, it seems reasonable to assume that overall or composite achievement probably falls somewhere in the mid-range with respect to the amount of within class, within school or within system

homogeneity. Undoubtedly some of the other variables, if measured as reliably as was the composite score, would show greater homogeneity and others less homogeneity than the composite score. One might expect variables such as capitalization skills or punctuation skills, which are learned primarily in the classroom, to show more within cluster homogeneity and variables such as vocabulary or reading to show less within cluster homogeneity.

Within school homogeneity would probably be less for pupils in junior and senior high school grades than for fourth graders, because such schools often draw pupils from larger more heterogeneous geographic areas. The within class homogeneity for junior and senior high school grades would probably vary depending on the type of class used. Homeroom classes, which are often constructed for administrative purposes only, might be less homogeneous than fourth grade classes. On the other hand, English classes, which would be more likely to contain pupils who were grouped by ability, might be more homogeneous than fourth grade classes.

With respect to states, Iowa has a relatively homogeneous population and the quality of schools is also quite homogeneous. Compared to the nation as a whole Iowa has a lower proportion of poor performers on the Iowa Tests of Basic Skills. Thus, classes, buildings and systems in Iowa would probably tend to be less homogeneous, i.e. have smaller intraclass correlation coefficients, than would those of many other states.

One further point should be made. Since the results of this investigation, together with the results of other investigations, indicate substantial homogeneity within natural clusters of pupils with respect to such characteristics as attitudes and behavior, answers to individual science test questions and overall achievement test scores, it is safe to assume that within cluster homogeneity will be a characteristic of many other variables of interest to the educational researcher. He should, therefore, anticipate such homogeneity and allow for it when planning research projects.

AERA, 1973 Annual Meeting
February, 1973

THE EFFECTS OF CLUSTER SAMPLING IN THE NORMING OF AN ACHIEVEMENT TEST BATTERY

by Jane Williams Bergsten

1. Population: ITBS composite test scores for fourth grade public school pupils in Iowa, January 1970 testing. Population consisted of 45,296 pupils in 1,794 classes in 987 buildings in 428 systems.

2. Size of Samples: 1/16 of population or about 2,800 pupils.

3. $$\frac{\text{error variance for cluster sample}}{\text{error variance for simple random sample}} = \text{DEFF} \doteq 1 + \rho\,(\overline{N} - 1)$$

where $\overline{N}$ = average number of pupils in a cluster
$\overline{N}$ = 106 pupils per system
$\overline{N}$ = 46 pupils per building
$\overline{N}$ = 25 pupils per class.

4.

Type of Cluster Sample	Grade-Equivalent Score: 30 or less	35 or less	45 or less	55 or less	60 or less
	$p \doteq .14$	$p \doteq .26$	$p \doteq .56$	$p \doteq .84$	$p \doteq .93$
	Design Effect (DEFF)				
Classes					
Unstratified	4.1	4.4	4.1	2.7	2.1
Stratified*	3.6	3.9	3.7	2.6	2.0
Buildings					
Unstratified	6.6	6.7	5.9	3.7	2.6
Stratified*	5.5	5.6	5.0	3.3	2.4
Systems					
Unstratified	56.3	55.4	45.2	24.4	12.6
Stratified*	30.1	29.4	27.2	14.8	7.3
Mixed**	8.3	8.9	8.5	5.2	3.6

5. Intraclass correlations:
(ρ's for systems) < (ρ's for buildings) < (ρ's for classes)

* Stratified by Size of System.
**Sampling buildings in the stratum containing the five largest systems and sampling systems in each of the remaining strata.